The incoherence condition in additive models

Sara van de Geer
Joint work with Lukas Meier and Peter Bühlmann
Seminar für Statistik, ETH Zürich

Berlin Workshop
Sparsity and Inverse Problems in Statistical Theory and Econometrics
December 6, 2008

Contents

1. Regression model

$$Y_i = f^0(X_i) + \epsilon_i, \ i = 1, \ldots, n,$$

with

$Y_i \in \mathbb{R}$,

$X_i \in \mathcal{X}$,

$f^0 : \mathcal{X} \to \mathbb{R}$ an unknown function.

We first look at the standard Lasso, and then study the

Additive model:

$\mathcal{X} = [0,1]^p$ with p large, and

$$f^0 \in \{f(x^{(1)}, \ldots, x^{(p)}) = \sum_{j=1}^{p} f_j(x^{(j)}) : \ f_j \in \mathcal{F}\}.$$

Notation

Let

$$Q_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

be the empirical distribution of the co-variables. Define

$$\| f \|_n^2 := \frac{1}{n} \sum_{i=1}^{n} f^2(X_i).$$

2. The Lasso

Let

$$f_\beta(\cdot) := \sum_{j=1}^{p} \beta_j \psi_j(\cdot) : \ \beta \in \mathbf{R}^p,$$

with $\{\psi_j\}$ a given dictionary of functions on $\mathcal{X}$.

Define the ℓ_1-norm

$$\|\beta\|_1 := \sum_{j=1}^{p} |\beta_j|.$$

Lasso:

$$\hat{\beta} := \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - f_\beta(x_i))^2 + \lambda \|\beta\|_1 \right\}.$$

Remark We assume the ψ_j are normalized:

$$\|\psi_j\|_n = 1, \ \forall \ j.$$

Equivalently, we could replace $\|\beta\|_1$ by

$$\sum_{j=1}^{p} \|\psi_j\|_n |\beta_j|,$$

in the definition of the Lasso.

3. Condition $\hat{\mathbb{C}}$

Define

$$S := \{j : \ \beta_j^0 \neq 0\}, s = |S|$$

$$\beta_S := \beta 1\{j \in S\}, \ \beta_{S^c} := 1\{j \notin S\},$$

Let

$$f = f_\beta := f_S + f_{S^c}$$

with

$$f_S := f_{\beta_S} = \sum_{j \in S} \beta_j \psi_j,$$

and

$$f_{S^c} := f_{\beta_{S^c}} = \sum_{j \notin S} \beta_j \psi_j.$$

Condition $\hat{\mathbb{C}}$: *For some constant $\hat{K}$ and for all β, satisfying*

$$(1-\eta)\|\beta_{S^c}\|_1 \leq (1+\eta)\|\beta_S\|_1,$$

we have

$$\|\beta_S\|_2 \leq \hat{K}\|f_\beta\|_n$$

Remark

Condition $\hat{\mathbb{C}}$ is called the *Compatibility Condition* in van de Geer (2007).

It is (roughly) the *Restricted Eigenvalue* (RE) Property in Bickel, Ritov and Tsybakov (2008).

The *Restricted Isometry Property* (RIP) (Candes and Tao (2007)) is sufficient.

Related: *Mutual Incoherence*, *Uniform Uncertainty Principle* (UUP), *Irrepresentability Condition*.

4. Oracle inequality

We define

$$(\epsilon, \psi_j)_n := \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \psi_j(x_i), \ j = 1, \ldots, p,$$

and

$$\mathcal{T}_0 := \{ \max_{1 \le j \le p} 2|(\epsilon, \psi_j)_n| \le \lambda_0 \}.$$

Lemma *Assume condition* $\hat{\mathrm{C}}$*, and that* $\lambda_0 \leq \eta\lambda$*. Then on* $\mathcal{T}_0$*, we have*

$$\|\hat{f} - f^0\|_n^2 + 2(1-\eta)\lambda\|\hat{\beta} - \beta^0\|_1 \leq 4\hat{K}^2\lambda^2 s.$$

Proof. See Bickel, Ritov and Tsybakov (2008).

5. Random design

We now no longer assume the ψ_j to be normalized for the empirical norm $\|\cdot\|_n$, i.e., we replace $\|\beta\|_1$ by $\sum_{j=1}^p \|\psi_j\|_n |\beta_j|$. Let

$$\hat{\Sigma} := \int \psi \psi^T dQ_n.$$

Let Σ be some symmetric $p \times p$ matrix. Define

$$\mathcal{T}_1 := \left\{ \max_{1 \leq j,k \leq p} \left| \frac{\hat{\sigma}_{j,k} - \sigma_{j,k}}{\sigma_j \sigma_k} \right| \leq \lambda_1 \right\}.$$

We now define the theoretical norm

$$\|f_\beta\|^2 = \beta^T \Sigma \beta.$$

To simplify the formula's, we assume the ψ_j are normalized for the theoretical norm $\|\cdot\|$, i.e., that $\sigma_j^2 := \|\psi_j\|^2 = 1$ for all j.

Example. Suppose that $X_1, \ldots, X_n$ are i.i.d. copies of X with distribution Q. Take

$$\Sigma = \int \psi \psi^T dQ.$$

Then for bounded or sub-exponential distributions, $\mathbf{P}(\mathcal{T}_1)$ is large for

$$\lambda_1 \sim \sqrt{\frac{\log p}{n}}.$$

The theoretical version of Condition $\hat{\mathrm{C}}$ is

Condition C *For a constant* K *all* β *satisfying*

$$(1-\eta)^2\|\beta_{S^c}\|_1 \leq (1+\eta)^2\|\beta_S\|$$

we have

$$\|\beta_S\|_2 \leq K\|f_\beta\|.$$

Lemma *Assume Condition C and the sparsity condition*

$$K^2\lambda_1 s \le \frac{3(1-\eta)^2}{16(1+\eta^2)}.$$

Then on $\mathcal{T}_1$, Condition $\hat{\mathrm{C}}$ holds, i.e., for all

$$(1-\eta)\sum_{j\notin S}\|\psi_j\|_n|\beta_j| \le (1+\eta)\sum_{j\in S}\|\psi_j\|_n|\beta_j|,$$

we have

$$\sum_{j\in S}\|\psi_j\|_n^2\beta_j^2 \le \hat{K}^2\|f_\beta\|_n^2,$$

where $\hat{K} = 2(1+\eta)K$.

Proof.

$$\|f\|_2 - \|f\|_n^2 \leq \lambda_1 \|\beta\|_1^2 \leq \lambda_1 \frac{2(1+\eta^2)}{(1-\eta)^2} \|\beta_S\|_1^2$$

$$\leq \lambda_1 \frac{2(1+\eta^2)s}{(1-\eta)^2} \|\beta_S\|_2^2 \leq \lambda_1 \frac{2K^2(1+\eta^2)s}{(1-\eta)^2} \|f\|^2$$

$$\leq \frac{3}{4} \|f\|^2.$$

So

$$\|f\|^2 \leq 4\|f\|_n^2.$$

If $\sum_{j \notin S} \|\psi_j\|_n |\beta_j| \leq \sum_{j \in S} \|\psi_j\|_n |\beta_j| (1+\eta)/(1-\eta)$, then

$$\|\beta_{S^c}\|_1 \leq \|\beta_S\| (1+\eta)^2/(1-\eta)^2.$$

So

$$\sum_{j \in S} \|\psi_j\|_n \beta_j^2 \leq (1+\eta)^2 \|\beta_S\|_2 \leq K^2 (1+\eta)^2 \|f\|^2 \leq 4K^2 (1+\eta) 2 \|f\|_n^2.$$

Corollary *Assume Condition C and the sparsity condition*

$$K^2\lambda_1 s \leq \frac{3(1-\eta)^2}{16(1+\eta^2)}.$$

Then for $\lambda_0 \leq \eta\lambda$*, we have on* $\mathcal{T}_0 \cap \mathcal{T}_1$*,*

$$\|\hat{f}-f^0\|_n^2+2(1-\eta)\lambda\|\hat{\beta}-\beta^0\|_1 \leq 16(1+\eta)^2K^2\lambda^2 s.$$

Moreover, on $\mathcal{T}_1$*,*

$$\|\hat{f}-f^0\|^2 \leq 4\|\hat{f}-f^0\|_n^2.$$

6. Additive model

Let $\mathcal{X} = [0,1]^p$ with p large. The standard Lasso considers the linear model

$$f_\beta(x^{(1)}, \ldots, x^{(p)}) = \sum_{j=1}^{p} \beta_j x^{(j)}.$$

We now generalize this to the nonparametric model

$$f(x^{(1)}, \ldots, x^{(p)}) = \sum_{j=1}^{p} f_j(x^{(j)}),$$

where $f_j \in \mathcal{F} := \{I(f_j) < \infty\}$,

with I the Sobolev norm

$$I^2(f_j) := \int |f_j^{(m)}(z)|^2 dz.$$

Uniting computational feasibility and oracle behavior

Let

$$\mathrm{pen}(f) := \sum_{j=1}^{p} \mathrm{pen}(f_j),$$

with

$$\mathrm{pen}(f_j) := \lambda^{\frac{2m}{2m+1}} \tau_n(f_j) + \lambda^{\frac{4m}{2m+1}} I^2(f_j),$$

and

$$\tau_n^2(f_j) := \|f_j\|_n^2 + \lambda^{\frac{4m}{2m+1}} I^2(f_j).$$

Nonparametric Lasso:

$$\hat{f} = \arg \min_{f=\sum_{j=1}^{p} f_j}$$

$$\left\{ \|Y - f\|_n^2 + \lambda^{\frac{2m}{2m+1}} \sum_{j=1}^{p} \tau_n(f_j) + \lambda^{\frac{4m}{2m+1}} \sum_{j=1}^{p} I^2(f_j) \right\}$$

7. Condition $\hat{\mathbb{C}}$

Define the *active* set

$$S := \{j : \|f_j^0\|_n \neq 0\},$$

and let $s := \mathrm{card}(S)$.

Condition $\hat{\mathbb{C}}$ *For all* f *satsifying*

$$(1-\eta)\sum_{j\notin S}\tau_n(f_j) \leq (1+\eta)\sum_{j\in S}\tau_n(f_j),$$

we have

$$\sum_{j\in S}\|f_j\|_n^2 \leq \hat{K}^2\left(\|f\|_n^2 + \lambda^{\frac{4m}{2m+1}}\sum_{j\in S}I^2(f_j)\right).$$

8. Oracle inequality

Empirical process

Let

$$\mathcal{T}_0 := \left\{ \sup_{f=\sum f_j} \frac{|2\sqrt{2}(\epsilon, \sum_{j=1}^p f_j)_n|}{\sum_{j=1}^p \tau_n(f_j)} \leq \lambda_0^{\frac{2m}{2m+1}} \right\}.$$

For sub-Gaussian errors, the choice

$$\lambda_0 \sim \sqrt{\log(p)/n},$$

gives the set $\mathcal{T}_0$ large probability.

Theorem *Suppose* $\lambda_0 \leq \eta\lambda$, *and that Condition* $\hat{\mathrm{C}}$ *is met. Then on the set* $\mathcal{T}_0$, *it holds that*

$$\|\hat{f} - f^0\|_n^2 + 2(1-\eta)\lambda^{\frac{2m}{2m+1}} \sum_{j=1}^{p} \tau_n(\hat{f}_j - f_j^0)$$

$$+\lambda^{\frac{4m}{2m+1}} \sum_{j=1}^{p} I^2(\hat{f}_j) \leq 3\lambda^{\frac{4m}{2m+1}} \sum_{j \in S} [I^2(f_j^0) + 8\hat{K}^2].$$

Remark One may take $\lambda \sim \sqrt{\log p/n}$. Thus $\lambda^{\frac{4m}{2m+1}}$ is of order $(\log p/n)^{\frac{2m}{2m+1}}$, which is up to the log-term the usual rate for estimating a m times differentiable function.
If f^0 has bounded smoothness $I(f_j^0)$ for all j, the rate is thus $s \times (\log p/n)^{\frac{2m}{2m+1}}$, with s being the number of active variables. This is, again up to the log-term, the same rate one would obtain if it was known beforehand which of the p functions are relevant.

9. Random design

Let $X_1, \ldots, X_n$ be i.i.d. copies of X. Let Q be the distribution of X and $\|\cdot\|$ be the $L_2(Q)$-norm. Let

$$\tau^2(f_j) := \|f_j\|^2 + \lambda^{\frac{4m}{2m+1}} I^2(f_j).$$

Condition C *For all f satisfying*

$$(1-\eta)^2 \sum_{j \notin S} \tau(f_j) \leq 4(1+\eta) \sum_{j \in S} \tau(f_j),$$

we have

$$\sum_{j \in S} \|f_j\|^2 \leq K^2 \left(\|f\|^2 + \lambda^{\frac{4m}{2m+1}} \sum_{j \in S} I^2(f_j) \right).$$

Let

$$\mathcal{T}_1 := \left\{ \sup_f \frac{|\|f\|_n^2 - \|f\|^2|}{(\sum_{j=1}^p \tau(f_j))^2} \leq \lambda_1 \right\}.$$

Lemma *Assume Condition C and the sparsity condition*

$$C_\eta K^2 \lambda_1^{\frac{2m-1}{2m+1}} s \leq 1.$$

Then on $\mathcal{T}_1$, Condition $\hat{\mathrm{C}}$ holds.

Corollary *Suppose Condition C and the sparsity condition. Then on* $\mathcal{T}_0 \cap \mathcal{T}_1$ *we have the oracle inequality for the nonparametric Lasso.*

Theorem *Suppose that for all* j*, the marginal distribution* Q_j *of* $X^{(j)}$ *has density* q_j *satisfying* $q_j \geq \eta_0 > 0$*. Then for* $\lambda_1 \sim \sqrt{\log p/n}$ *the set* $\mathcal{T}_1$ *has large probability.*

10. On Condition C

Well-conditioned active set condition *We say that the active set S is well conditioned if for some constant $0 < \psi_0 \leq 1$, and for all $\{f_j\}_{j \in S}$,*

$$\sum_{j \in S} \|f_j\|^2 \leq \|\sum_{j \in S} f_j\|^2 / \psi_0^2.$$

The inner product between functions f and $\tilde{f}$ is denoted by $(f, \tilde{f}) := \int f \tilde{f} dQ$.

No perfect canonical dependence condition *We say that the active and non-active variables have no perfect canonical dependence, if for a constant* $0 \leq \rho_0 < 1$*, and all* $\{f_j\}_{j=1}^p$*, we have for* $f_S := \sum_{j \in S} f_j$ *and* $f_{S^c} := \sum_{j \notin S} f_j$*, that*

$$\frac{|(f_S, f_{S^c})|}{\|f_S\| \|f_{S^c}\|} \leq \rho_0.$$

The next Lemma makes the link between the Condition C and the above two conditions.

Lemma *Let* $f = f_S + f_{S^c}$ *satisfy*

$$\frac{|(f_S, f_{S^c})|}{\|f_S\|\|f_{S^c}\|} \leq \rho_0 < 1.$$

Then

$$\|f_S\|^2 \leq \|f\|^2/(1-\rho_0^2).$$

Corollary *A well-conditioned active set in combination with no perfect canonical dependence, implies the Condition C with* $K^2 = \frac{1}{\psi_0^2(1-\rho_0^2)}$.

Lemma *Let q be the density of X, q_S be the density of $\{X^{(j)}\}_{j\in S}$ and q_{S^c} the density of $\{X^{(j)}\}_{j\notin S}$. Suppose*

$$\int \frac{q}{q_S q_{S^c}} \leq 1 + \rho_0^2.$$

Then the no perfect canonical dependence condition holds.

References

Bickel, P. Ritov, Ya. and Tsybakov A. (2008). Simultaneous analysis of Lasso and Dantzig selector. To appear in Annals of Statistics.

Bunea, F. and Tsybakov, A.B. and Wegkamp, M.H. (2007), Sparsity oracle inequalities for the Lasso, *Electr. Journal of Statist.* **1**

Candes, E. and Tao, T. (2007) The Dantzig selector: statistical estimation when p is much larger than n, *Ann.Statist.*

B. Tarigan and S.A. van de Geer (2006). Classifiers of support vector machine type, with ℓ_1 penalty. *Bernoulli* **12**, 1045–1076.

van de Geer, S. (2007). The deterministic Lasso. JSM proceedings, paper nr. 489.

van de Geer, S. (2008). High-dimensional generalized linear models and the Lasso. *Ann. Statist.*

Meier, L. Bühlmann, P. van de Geer, S. (2008). High-dimensional additive modeling. Techn. Report SfS

THANKS!